CONGESTION AVOIDANCE WITHIN AGGREGATE CHANNELS

FIELD OF THE INVENTION

This invention relates generally to avoiding congestion in networks such as

5   Internet Protocol (IP)-based networks, and more particularly to avoiding congestion in

such networks within aggregate channels.

BACKGROUND OF THE INVENTION

Computerized networks, such as the Internet, have become increasingly popular.

10   Such networks commonly allow computers at geographically distant locations to

exchange information. For example, a computer user in the United Kingdom may be able

to download software from a server located in the United States.

A disadvantage that has become apparent with such networks is, however,

increased congestion, slowing down the speed at which data can be communicated at

15   particular times. For example, a computer user downloading a multimedia video clip

may experience choppy playback, or playback stoppage in the middle of a clip, due to the

network (e.g., the Internet) being congested to the point where all the data transmitted

thereover cannot be transmitted in a timely manner. As more and more data-intensive

applications have become commonplace on the Internet, such as multimedia video

20   streaming, congestion has only gotten worse, and does not promise to let up in the future.

Explicit Congestion Notification (ECN) has been proposed as a manner by which

to implement congestion avoidance, as described in K. Ramakrishnan and S. Floyd, A

proposal to add Explicit Congestion Notification (ECN) to IP, Request for Comments

RFC 2481, Internet Engineering Task Force, January 1999, which is hereby incorporated

by reference. However, as has been proposed within this reference, ECN is very particular to TCP. This means that usage of a great amount of network bandwidth that relies on UDP or other non-TCP protocols is unaffected by ECN, such that congestion avoidance strategies cannot be triggered by ECN. Further, ECN as has been proposed

5   within this reference relies on notifying only those users (e.g., channels) that are causing congestion, instead of all related users (or, channels), restricting the manner by which congestion can be alleviated.

For these and other reasons, there is a need for the present invention.

10   **SUMMARY OF THE INVENTION**

The invention relates to congestion avoidance within aggregate network channels. In one embodiment, a method first transmits a packet associated with a first channel of a plurality of related channels from a source protocol layer (e.g., a source IP layer) of a source through a network (e.g., the Internet). Next, the method triggers an ECN event by

15   the packet at the network. Finally, at least one channel is determined to have decreased packets transmitted therethrough, in response to the triggering of the ECN event (e.g., based on a congestion pricing criteria).

In one particular embodiment, the method includes determining at a computer program of the source or a higher layer than the source IP layer (e.g., a UDP, TCP, or

20   other layer) a selected channel or channels of the plurality of related channels to have reduced packets transmitted therethrough, in response the ECN event. In one particular embodiment, the ECN event is based on a congestion pricing criteria. Thus, in one

2

particular embodiment, the source proportionally reduces the packets transmitted through the selected channel(s) according to a congestion pricing criteria.

As described, embodiments of the invention have advantages over the prior art. The ECN according to an embodiment of the invention is able to deal with all different higher layers, such as end user applications (either of the source or destination), as well as TCP, UDP, etc., and is not specific or limited to one particular type of higher layer, such as TCP.

Furthermore, in the embodiment of the invention where there are a plurality of related channels, in effect all these channels become aware of the congestion on any one thereof -- that is, even if the congestion is occurring on a particular one of the channels, the situation can be alleviated by decreasing traffic on a different one or more of the channels. This means if a relatively important channel is causing the congestion, congestion can be alleviated not by decreasing transmission through this channel, but rather by decreasing transmission through another, less important, related channel.

The invention includes methods, systems, computer-readable media, and computers of varying scope. Still other aspects, embodiments, and advantages of the invention will become apparent by reference to the drawings and by review of the detailed description that follows.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a diagram of an operating environment in conjunction with which embodiments of the invention can be practiced;

FIG. 2 is a diagram of a system according to an embodiment of the invention;

3

FIG. 3 is a diagram of an aggregate of channels, according to an embodiment of the invention;

FIG. 4 is a flowchart of a method according to an embodiment of the invention; and,

5   FIG. 5 is a diagram of a computer according to one embodiment of the invention.

DETAILED DESCRIPTION OF THE INVENTION

In the following detailed description of exemplary embodiments of the invention, reference is made to the accompanying drawings which form a part hereof, and in which

10   is shown by way of illustration specific exemplary embodiments in which the invention may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the invention, and it is to be understood that other embodiments may be utilized and that logical, mechanical, electrical and other changes may be made without departing from the spirit or scope of the present invention. The

15   following detailed description is, therefore, not to be taken in a limiting sense, and the scope of the present invention is defined only by the appended claims.

Some portions of the detailed descriptions which follow are presented in terms of algorithms and symbolic representations of operations on data bits within a computer memory. These algorithmic descriptions and representations are the means used by those

20   skilled in the data processing arts to most effectively convey the substance of their work to others skilled in the art. An algorithm is here, and generally, conceived to be a self-consistent sequence of steps leading to a desired result. The steps are those requiring physical manipulations of physical quantities. Usually, though not necessarily, these

quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated.

It has proven convenient at times, principally for reasons of common usage, to refer to these signals as bits, values, elements, symbols, characters, terms, numbers, or the like. It should be borne in mind, however, that all of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities. Unless specifically stated otherwise as apparent from the following discussions, it is appreciated that throughout the present invention, discussions utilizing terms such as processing or computing or calculating or determining or displaying or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system's registers and memories into other data similarly represented as physical quantities within the computer system memories or registers or other such information storage, transmission or display devices.

The detailed description is divided into sections as follows. First, a description of an operating environment in conjunction with which embodiments of the invention may be practiced is provided. Next, terms and topics central to understanding embodiments of the invention are described. Following this, systems according to varying embodiment of the invention are described, relying on the terms and topics described in the previous section of the detailed description, and including a description of a specific computer embodiment of the invention. Then, methods according to varying embodiments of the invention are presented. Finally, a conclusion is provided.

Operating Environment

Referring to FIG. 1, a diagram of the hardware and operating environment in conjunction with which embodiments of the invention may be practiced is shown. The description of FIG. 1 is intended to provide a brief, general description of suitable

5 computer hardware and a suitable computing environment in conjunction with which the invention may be implemented. Although not required, the invention is described in the general context of computer-executable instructions, such as program modules, being executed by a computer, such as a personal computer. Generally, program modules include routines, programs, objects, components, data structures, etc., that perform

10 particular tasks or implement particular abstract data types.

Moreover, those skilled in the art will appreciate that the invention may be practiced with other computer system configurations, including hand-held devices, multiprocessor systems, microprocessor-based or programmable consumer electronics, network PC's, minicomputers, mainframe computers, and the like. The invention may

15 also be practiced in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

The exemplary hardware and operating environment of FIG. 1 for implementing

20 the invention includes a general purpose computing device in the form of a computer 20, including a processing unit 21, a system memory 22, and a system bus 23 that operatively couples various system components include the system memory to the processing unit 21. There may be only one or there may be more than one processing unit 21, such that the

processor of computer 20 comprises a single central-processing unit (CPU), or a plurality

of processing units, commonly referred to as a parallel processing environment. The

computer 20 may be a conventional computer, a distributed computer, or any other type

of computer; the invention is not so limited.

5   The system bus 23 may be any of several types of bus structures including a

memory bus or memory controller, a peripheral bus, and a local bus using any of a

variety of bus architectures. The system memory may also be referred to as simply the

memory, and includes read only memory (ROM) 24 and random access memory (RAM)

25. A basic input/output system (BIOS) 26, containing the basic routines that help to

10   transfer information between elements within the computer 20, such as during start-up, is

stored in ROM 24. The computer 20 further includes a hard disk drive 27 for reading

from and writing to a hard disk, not shown, a magnetic disk drive 28 for reading from or

writing to a removable magnetic disk 29, and an optical disk drive 30 for reading from or

writing to a removable optical disk 31 such as a CD ROM or other optical media.

15   The hard disk drive 27, magnetic disk drive 28, and optical disk drive 30 are

connected to the system bus 23 by a hard disk drive interface 32, a magnetic disk drive

interface 33, and an optical disk drive interface 34, respectively. The drives and their

associated computer-readable media provide nonvolatile storage of computer-readable

instructions, data structures, program modules and other data for the computer 20. It

20   should be appreciated by those skilled in the art that any type of computer-readable media

which can store data that is accessible by a computer, such as magnetic cassettes, flash

memory cards, digital video disks, Bernoulli cartridges, random access memories

(RAMs), read only memories (ROMs), and the like, may be used in the exemplary operating environment.

A number of program modules may be stored on the hard disk, magnetic disk 29, optical disk 31, ROM 24, or RAM 25, including an operating system 35, one or more application programs 36, other program modules 37, and program data 38. A user may enter commands and information into the personal computer 20 through input devices such as a keyboard 40 and pointing device 42. Other input devices (not shown) may include a microphone, joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 21 through a serial port interface 46 that is coupled to the system bus, but may be connected by other interfaces, such as a parallel port, game port, or a universal serial bus (USB). A monitor 47 or other type of display device is also connected to the system bus 23 via an interface, such as a video adapter 48. In addition to the monitor, computers typically include other peripheral output devices (not shown), such as speakers and printers.

The computer 20 may operate in a networked environment using logical connections to one or more remote computers, such as remote computer 49. These logical connections are achieved by a communication device coupled to or a part of the computer 20; the invention is not limited to a particular type of communications device. The remote computer 49 may be another computer, a server, a router, a network PC, a client, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer 20, although only a memory storage device 50 has been illustrated in FIG. 1. The logical connections depicted in FIG. 1 include a local-area network (LAN) 51 and a wide-area network (WAN) 52. Such

networking environments are commonplace in office networks, enterprise-wide computer

networks, intranets and the Internal, which are all types of networks.

When used in a LAN-networking environment, the computer 20 is connected to

the local network 51 through a network interface or adapter 53, which is one type of

5   communications device. When used in a WAN-networking environment, the computer

20 typically includes a modem 54, a type of communications device, or any other type of

communications device for establishing communications over the wide area network 52,

such as the Internal. The modem 54, which may be internal or external, is connected to

the system bus 23 via the serial port interface 46. In a networked environment, program

10   modules depicted relative to the personal computer 20, or portions thereof, may be stored

in the remote memory storage device. It is appreciated that the network connections

shown are exemplary and other means of and communications devices for establishing a

communications link between the computers may be used.

15   **Systems**

In this section of the detailed description, systems according to varying

embodiments of the invention are described. This section utilizes terms and topics that

have been defined and explained in the previous section of the detailed description;

therefore, for understanding of the systems described, reference should also be made to

20   the previous section of the detailed description.

Referring first to FIG. 2, a diagram of a system according to an embodiment of

the invention is shown. A source 200 sends packets of data, such as the packet 202, to a

destination 204, via a network layer 206. It is noted that the term network layer, as used

throughout the application, is not to be confused with the strict sense of a network layer

as used in the context of protocols known within the art (e.g., the "network protocol

layer"). Rather, the term as used herein extends to the mechanism that communicatively

couples the source 200 to the destination 204, such as a network (e.g., the Internet), as

5   those of ordinary skill within the art can appreciate.

Each of the source 200 and the destination 204 can be a computer -- for example,

in one embodiment, a computer such as has been described already in conjunction with

FIG. 1 (e.g., including a processor and computer-readable media such as memory and a

hard disk drive). Furthermore, each of the source 200 and the destination 204 include a

10   protocol layer (layers 210 and 212, respectively), and one or more higher layers (such as

layers 214 and 216, respectively, and which can also be protocol layers, as those of

ordinary skill within the art can appreciate). The invention itself is not so limited,

however. The source 200 is referred to as the source, and the destination 204 is referred

to as the destination, because normally the direction of data packet transfer between the

15   two is from the former to the latter; however, the destination 204 is still able to send

packets of data to the source 200. The nomenclature is thus just indicative of which of

the two is typically the source of data, and which is typically the destination of data.

The term layer generally refers to a different component of a networking system,

such that each layer has assigned responsibilities and passes through data packets to other

20   layers. For example, within a TCP/IP system, a protocol layer is the IP layer, while

higher layers include the TCP layer, an application layer, as well as potentially other

layers. Thus, an application layer may generate the data to be sent to a destination, while

the TCP layer adds control, error-checking and other such information, and the IP layer

10

packetizes the data and actually sends it over the network layer (for example, the Internet). When received by the destination, the IP layer depacketizes the data, the TCP layer ensures that the data was received correctly, such that it is finally received by an application layer. However, the invention is not so limited to a particular or any layered

5   networking scheme.

As has been stated, the network layer 206 in one embodiment is the Internet, as known in the art; it can in other embodiments also be an intranet, extranet, virtual private network (VPN), any other type of network, or any combination of such networks -- the invention is not so limited. As shown in FIG. 2, the network layer 206 includes an

10   aggregate of related channels, including channels 218 and 220. The channels 218 and 220 are referred to as related according to a predetermined criteria, such that the channels connect the source 200 to the destination 204 via the same path through the network layer 206. For example, the channels may be related because they transmit the audio and video, respectively, of a teleconference, from the source 200 to the destination 204.

15   It is noted in particular that the channel 218 has a congestion point 222. This means that the channel 218 cannot accommodate the same transmission rate of packets therethrough as normal, or as compared with the channel 220, for instance. This may result in the delay of packets sent by the source 200 from reaching the destination 204, or in one or more data packets being lost. The congestion point 222 can result from too

20   many packets being sent by the source 200 as compared to the advertised rate of the channel 218, for instance, or as a result of overloading of the network layer 206 in general, among other causes.

In one embodiment of the invention, an explicit congestion notice (ECN) is

generated as a result of the congestion within a channel, such as the channel 218 as

shown in FIG. 2. (That is, the ECN is triggered as a result of congestion generally, and

specifically as a result of a packet of data transmitted through the channel causing the

5   congestion.) ECN refers to notification in response to actual or predicted congestion

within a network, such as the Internet, as caused by a given packet of data through a

channel (such as an IP channel). That is, an ECN may result from actual congestion, or

from the network determining that congestion will occur soon -- for example, as

determined in conjunction with the congestion pricing criteria, as described in a further

10   section of the detailed description. The aim of the ECN is typically to encourage or

mandate users to cooperate and so run the network in a "sensible" state, namely a mostly

uncongested one.

The ECN can be generated, for instance, by the network layer 206, which notices

that the channel 218 has become congested, although the invention is not so limited.

15   Ultimately, however, a policy mechanism 224, based on the ECN, determines at least one

channel of the aggregate of related channels to have decreased transmission of packets

therethrough, to reduce congestion on the congested channel. Thus, in the embodiment

of FIG. 2, the policy mechanism 224 can reduce packet transmission of the congested

channel 218, or, significantly, can also reduce packet transmission of the non-congested

20   channel 220, which frees up more bandwidth for the channel 218.

The invention is not particularly limited as to the manner by which the policy

mechanism 224 receives notice of the ECN, nor as to the criteria by which it alleviates

network congestion. Thus, as shown in FIG. 2, dotted lines extend from the policy

12

mechanism 224 to each of the source 200, the destination 204, and the network layer 206, to indicate that the policy mechanism is able to reside in any of them. Furthermore, in one embodiment, the policy mechanism 224 is a computer program (e.g., a congestion policy program executed by the processor from the medium of the program), although

5   again the invention is not so limited.

In the embodiment where the policy mechanism 224 resides within the source 200, feedback of the ECN to the source 200 (so that the mechanism 224 is aware that the ECN has been generated) is achieved as follows. A data packet sent by the source 200 triggering the ECN is marked to indicate the ECN has been generated. For example, one

10 bit of the data packet may be switched on to indicate that the ECN has been generated (or, more than one data bit, depending on how much information is desired to be transferred about the congestion -- e.g., where specifically within the network the congestion occurred, etc.).

The manner by which marking of data packets to indicate that the ECN has been

15 generated is not limited by the invention, however. In one embodiment, a counter of the number of packets that should be marked is kept running, and as soon as a loss event occurs, marking is started from there until the right number has been marked (marking in this context thus refers to marking packets with ECN). In another embodiment, a threshold is set, probabilistically marking packets in the queue once this threshold is

20 exceeded. In still another embodiment, a virtual queue is run, with reduced rate and capacity, and packets are marked accordingly.

Once the data packet has reached the destination 204, the destination 204 sends a return packet also having its ECN bit turned on, so that upon receipt at the source 200,

the policy mechanism 224 is aware that the ECN has been generated. However, the

invention is not so limited -- the feedback of the ECN can occur in other manners, as

well. Thus, when the policy mechanism 224, in the embodiment where it resides within

the source 200, learns of the ECN, it must then determine which of the aggregate of

5   related channels to decrease throughput on. In one embodiment, this is performed in

conjunction with the congestion pricing criteria as is been described elsewhere in this

application -- that is, the congestion pricing criteria determines the reaction made by the

mechanism 224: which channel should have reduced bandwidth, and by how much.

In another embodiment of the invention the policy mechanism 224 resides within

10   the destination 204. In such a situation, for example, the mechanism 224 may receive

knowledge of the ECN by the packet of data sent by the source 200 that triggered the

ECN having a marked bit, such that the data packet is ultimately received by the

mechanism 224. The mechanism 224 then is able to determine which of the aggregate of

channels is to have reduced throughput, such that this information is conveyed to the

15   source 200 (e.g., via a data packet sent by the destination 204, or by another mode of

communication, etc.), so that the source 200 knows which channel to throttle back on.

Once this determination is made, the mechanism 224 then causes the reduced throughput

based on its recommendation, in a manner not limited by the invention.

In another embodiment, the policy mechanism 224 resides within the network

20   layer 206. Thus, as soon as an ECN is generated by the network layer 206, the

mechanism 224 receives notice thereof, and is able to determine which channel should

have reduced throughput. Again, however, the mechanism 224 must communicate this

14

information to the source 200, so that the source can implement the policy to reduce

throughput on the selected channel.

For further understanding of embodiments of the invention, reference is now

made to FIG. 3, which is a diagram showing an aggregate of three related channels.

5   Specifically, within a network layer 300 are channels 302, 304 and 306, where 302 is

used by a source for communication of video information of a teleconference transmitted

by the UDP protocol known within the art, 304 is used by the source for communication

of audio information of the teleconference also transmitted by the UDP protocol, and 306

is used by the source for communication of mouse click information of the teleconference

10   (i.e., assuming a computer teleconference) transmitted by the TCP protocol known within

the art.

Thus, if an ECN is generated on channel 302, indicating that this channel is

congested (perhaps, for example, as a result of the high bandwidth required by video

data), then the policy mechanism -- not shown in FIG. 3 -- must determine which of the

15   channels 302, 304 and 306 to reduced packet transmission on in order to decrease

network congestion. For example, based on the criteria used by the mechanism, the

decision may be made to decrease the video data throughput of itself (since this would

cause only a degradation in the quality of the video data, and not a complete loss of the

signal), or decrease the audio data throughput, or the mouse data throughput -- even

20   though the latter two did not contribute to the network congestion.

Referring next to FIG. 5, a diagram of a computer according to an embodiment of

the invention is shown. The computer includes a protocol layer 500, a congestion policy

computer program 502 (which can be executed by a processor of the computer from a

15

computer-readable medium such as a memory thereof, both of which are not shown in

FIG. 5), a plurality of queues (e.g., buffers) such as queue 504, and a corresponding

plurality of filters such as filter 506. The latter each couple to a different one of an

aggregate of related channels, such as channel 508, which connect with and through a

5   network or network layer not shown in FIG. 5. furthermore, the policy program 502 is

responsive to an ECN 510.

The protocol layer 500 is the layer that packetizes data for transmission through

the network, and can be either a destination protocol layer of a destination computer, or a

source protocol layer of a source computer (that is, the computer of FIG. 5 can be either a

10   source or a destination computer). It can be said to include the plurality of related

channels, such as the channel 508. More specifically, access from the protocol layer 500

to the plurality of related channels is controlled by the plurality of queues, such as the

queue 504, and the plurality of filters, such as the filter 506.

Each queue has a modifiable depth, such that in one embodiment a new data

15   packet cannot enter a queue until there is an empty space. Controlling queue depth, thus,

is one manner by which a channel can be throttled back -- reducing its queue depth

causes the channel to be throttled back. Likewise, each filter is able in one embodiment

of the invention to specify the specific rate at which data packets leave the queue to enter

the channel, as well as in one embodiment the types of data packets that are so allowed to

20   enter the channel. Thus, controlling the criteria of a filter as to when and which packets it

allows to proceed to the channel is another manner by which a channel can be throttled

back.

The congestion policy program 502 determines at least one channel of the plurality of related channels to have decreased transmission of packets therethrough, such as at least one channel other than a congested channel, in response to the ECN 510. That is, assuming that one of the plurality of related channels, such as the channel 508,

5   becomes congested as a result of a data packet being sent (in particular), the network generates an ECN 510, feedback regarding which is received by the program 502. The program 502 can determine which channel or channels to reduce transmission through, and how much to reduce transmission, via any criteria, such as a congestion pricing criteria. Moreover, the program 502 can in one embodiment implement this

10   determination by modifying the filters and the queues of the affected channels, so that the channel only transmits the desired amount of data packets.

Method

A computer-implemented method according to an embodiment of the invention is

15   shown in FIG. 4. The computer-implemented method is desirably realized at least in part as one or more programs running on a computer -- that is, as a program executed from a computer-readable medium such as a memory by a processor of a computer. The programs are desirably storable on a machine-readable medium such as a floppy disk or a CD-ROM, for distribution and installation and execution on another computer.

20   Referring now to FIG. 4, in 400, a data packet associated with a first channel of an aggregate of related channels is transmitted, from a source protocol layer of a source, through a network. As has been described, the channels within the aggregate are related in that they have a common source and a common destination. The source protocol layer

17

can be one of many different layers within the source, which can also including such

higher level layers as end-user applications, and intermediate control layers, such as a

TCP layer as known within the art. The source in one embodiment is a computer, for

example, as has been described in conjunction with FIG. 1 already. The source protocol

5   layer is an IP protocol layer, as has been described. Furthermore, in one embodiment, the

network comprises the Internet, a virtual private network, an intranet, an extranet, or a

combination of two or more of these, although again the invention is not so limited.

In 402, an ECN event is triggered by the packet at the network. The packet, in

being transmitted from the source protocol layer of the source and through the network, is

10   thus congesting the network, causing the network in response to indicate this by

generating an ECN event. The manner by which the ECN event is triggered is also not

limited by the invention. Thus, the ECN event can be triggered in accordance with a

congestion pricing criteria, as is described in the next section of the detailed description,

and can result in a marking of one of at least one bit of the data packet that triggered the

15   ECN, by the network.

In 404, feedback is received of the ECN event. The feedback is received at one or

more of the source; a destination having a destination protocol layer and one or more

upper layers (in other words, comparable to the source) that is the intended receiver of

packets of data sent by the source; and the network. Thus, the generation of the ECN

20   event in 402 does not necessarily mean in some embodiments of the invention that the

source, destination and/or network is able to be aware of the ECN. Therefore, in 404, the

desired one or more of these networking system components receives feedback as a

manner by which they become aware of the ECN.

18

Furthermore, the feedback may be received at any layer of the destination or source -- e.g., the protocol layer, a higher layer such as a control layer like TCP, or an even still higher layer such as an end-user application layer. In general, the layer that receives feedback is the location where the policy mechanism, described in the previous

5   section of the detailed description, resides. The feedback thus may be received at a computer program of a higher layer of the source or the destination.

Where the source is to receive feedback of the ECN, it is noted that in at least one embodiment of the invention, the destination is to provide this feedback to the source. That is, the packet marked with the ECN and sent by the source is received at the

10   destination. Thus, the destination then communicates its receipt of a packet marked with the ECN. This can be done in one embodiment by sending back a data packet to the source (specifically, from a protocol layer of the destination, herein referred to as the destination protocol layer), to indicate that it received a packet marked with the ECN, or communicating through another manner.

15   Finally, in 406, at least one channel is determined to have decreased packets transmitted therethrough, in response to the triggering of the ECN event. This is pursuant to as has been described in the previous section of the detailed description. Thus, to reduce congestion, at least one of the aggregate of channels must have its throughput decreased -- either including or not including the channel through which the packet

20   triggering the ECN. In one embodiment, the at least one channel is determined in accordance with a congestion pricing criteria, as has been described, although the invention is not so limited.

In one embodiment, the determination is made by a policy mechanism, as has

been described, which resides at one of the source, the destination and the network.

Thus, the determination is made by one of the source, the destination and the network as

well. In the case of the destination and the source, the determination may be made at any

5   layer thereof, including the protocol layer (e.g., the IP layer), the control layer (e.g., the

TCP layer), and any other higher layer, such as an end-user application layer. The

invention is not so limited.

Congestion Pricing Criteria

10   At least some embodiments of the invention utilize a specific congestion pricing

criteria, now described, to achieve reduced throughput on one or more of a set of related

channels, to avoid, prevent, or alleviate network congestion. That is, the congestion

pricing criteria described is one manner by which such network congestion can be

avoided, prevented, or alleviated.

15   The basic framework of a congestion pricing criteria according to one

embodiment of the invention has two primary elements. First, the network supplies

feedback to the users (which can be end-user applications, a higher layer, a protocol

layer, etc.), which reflects the cost of congestion (more precisely, the shadow price of

congestion). Second, the users incur some cost associated with their actions, but

20   otherwise can react as they please.

It is assumed that the cost of congestion is the number of lost packets. Then, all

packets contributing to a loss should be marked, such as by an ECN mark, to reflect the

20

shadow price of congestion. In other words, if there is a load y on a resource, incurring

cost at rate $C(y)$, the congestion price of shadow price is the derivative

$$p(y) = \frac{d}{dy} C(y).$$

If the load y is a random variable and the cost is taken to be the expected rate of loss, then

the marking strategy in a single buffer returns the appropriate derivative information,

scaled by the load.

Furthermore, it is supposed that a finite number of users R (indexed by r) use a

finite set of resources J (indexed by j), where the 0-1 incidence matrix A_{jr} indicates

whether user r uses resource j or not. An elastic user's preference for the amount of

bandwidth x_r is characterized by a concave utility function $U_r(x)$. Thus a user would seek

to maximize the total net utility, maximizing

$$\sum_r U_r(x_r) - \sum_j C_j \left(\sum_r A_{jr} x_r \right),$$

over $x_r \geq 0$; $r \in R$. This optimization problem has the solution, for non-zero x_r,

$$U_r'(x_r) \equiv \frac{d}{dx_r} U_r(x_r) = \sum_{j \in r} p_j(y_j)$$

where $j \in r$ is the set $\{j : A_{jr} = 1\}$, with a corresponding load on resource j given by

$$y_j = \sum_r A_{jr} x_r.$$

That is, at the social optimum, the derivative of a user's utility function exactly matches

the sum of the shadow prices of all resources along the user's route.

Usually, the network has no knowledge about the users' utility functions.

However, if a user is charged at a rate proportional to the amount of bandwidth x_r

received, for example, at a rate $t_r x_r$, users will seek to maximize their net return, that is, they will try to maximize

$$U_r(x_r) - t_r x_r$$

over $x_r \geq 0$. Thus, if the prices are right, namely if

5

$$t_r = \sum_{j \in r} p_j(y_j)$$

and if each user acts to maximize his or her own net benefit, then they will at the same time evolve towards the system of social optimum. As the network load changes, the network may need to update its prices.

Therefore, the feedback signal $x_r \sum_{j \in r} p_j(y_j)$ is used, where feedback for a stream

10 should be proportional to the stream and also proportional to the shadow prices (congestion costs) along the route.

If each user r uses the updating equation

$$\frac{d}{dt} x_r(t) = \kappa_r \left(x_r(t) U_r'(x_r(t)) - x_r(t) \sum_{j \in r} p_j(y_j(t)) \right)$$

for some positive gain parameter κ_r, where $y_j(t) = \sum_{j \in r} x_r(t)$, then, provided the utility

15 function is strictly concave and the cost function is strictly convex, defined over $x \geq 0$, solutions will converge to the unique equilibrium.

This convergence is specifically now proved. A function is first defined,

$$\Psi(x) = \sum_j U_r(x_r) - \sum_j C_j \left(\sum_{j \in s} x_s(t) \right).$$

Under the concave and convex assumptions on the functions U and C, respectively, Ψ is

20 a Lyapunov function for the differential equation

22

$$\frac{d}{dt}x_r(t) = \kappa_r\left(x_r(t)U'_r(x_r(t)) - x_r(t)\sum_{j\epsilon r}p_j(y_j(t))\right)$$

in which all trajectories converge to the vector x maximizing this function. This is unique

for strictly concave U and strictly convex C, and interior to the region. Next, consider

$$\frac{\partial}{\partial x_r}\Psi(x) = U'_r(x_r(t)) - \sum_{j\epsilon r}p_j(y_j(t));$$

5   setting these derivatives to zero yields the maximum. Differentiating with respect to t

gives

$$\frac{d}{dt}\psi(x(t)) = \sum\kappa_r x_r(t)\left(U'_r(x_r(t)) - \sum_{j\epsilon r}p_j(y_j(t))\right)^2$$

which is strictly positive except when x is the equilibrium value -- hence Ψ is a Lyapunov

function, and the result proved.

10   It is also noted that for utility functions of the form

$$U_r(x_r) = w_r\log x_r$$

for some willingness-to-pay constant w_r, at the optimum,

$$w_r = x_r\sum_{j\epsilon r}p_j(y_j).$$

The resulting allocation is proportionally fair, which is related to a Nash arbitration

15   scheme, as known within the art.

Next, it is shown how such an allocation can be applied to arbitrary utility

functions if w_r is treated as input to the user's control algorithm in place of true

preferences. First, it is supposed that each user r uses the updating equation

$$\frac{d}{dt}x_r(t) = \kappa_r\left(w_r(t) - x_r(t)\sum_{j\epsilon r}p_j(y_j(t))\right)$$

23

and updates w_r according to

$$w_r(t) = x_r(t)U'_r(x_r(t)).$$

In this case, too, the solution converges to the global optimum.

It is noted that this holds true even if instead a user is mandated to use a certain

5   updating function, subject to certain conditions. This can be the case if a user or end-

system has a restricted choice of flow-control options, for example.

This case of mandated controls is now proved. The updating equation is

considered for some fixed concave functions F_r. It is supposed the derivative of F_r has a

continuous inverse, and that there are some proxy variables w_r with associated functions

10   g_r satisfying

$$F'_r(x_r) = p_r$$

$$x_r = w_r g_r(p_r).$$

Therefore, there are vectors $w = (w_r, r \in R), x = (x_r, r \in R), p = (p_r, r \in R)$ such that if w_r

solves the user problem, which maximizes

15    $U_r(w_r g_r(p_r)) - p_r w_r g_r(p_r)$ over $w_r \geq 0$,

then x also solves the system optimum. In this case the p_r are the sum of the shadow

prices along a route r.

Thus, in the case when the optimum lies in the interior of a region, the equation

$$\frac{d}{dt}x_r(t) = \kappa_r \left(w_r(t) - x_r(t) \sum_{j \in r} p_j(y_j(t)) \right)$$

20   converges to the unique equilibrium satisfying

$$F'_r(x_r) = \sum_{j \in r} p_j(y_j)$$

but at the user optimum

$$U_r'(x_r) = p_r.$$

Therefore, using

$$F_r'(x_r) = p_r$$

5
$$x_r = w_r g_r(p_r),$$

this is a solution to the problem previously described as

$$U_r'(x_r) \equiv \frac{d}{dx_r} U_r(x_r) = \sum_{j \epsilon r} p_j(y_j), \text{ completing the proof.}$$

Next, the case where a user has a number of streams that are related, which is
equivalent to the case where the "users" are not independent, is considered. This can
10   model a multimedia connection, for example. The case where a user r may have a utility

function $U_r(\vec{x}_r)$ that is a function of several variables is considered. Provided that this is
a concave function, the user optimum still converges to the system optimum, provided
that downhill steps are taken, but now the user can choose the extra information to alter
the adaptation. For instance, steepest descent or gradient projection methods, as known
15   within the art, can be used to change the way the user reacts, which equate to trading off
marks between streams, and altering rates of convergence.

This case of correlated feedback is now more specifically considered. Two
related streams are considered, and for simplicity it is assumed that they use the same set
of resources. The component feedback signals are then

20
$$x_i \frac{\partial U_r(x_i)}{\partial x_i} - x_i \sum_{j \epsilon r} p_j(y_j) \quad i = 1, 2,$$

25

which can be scaled in any manner to update the vector (x_1, x_2). It is noted that this is

equivalent to changing κ_i. Marks can also be traded among streams, in one embodiment,

using correlated information to increase reaction speed. For example, a proportion of

stream 2 marks to reduce stream 1 can be used, provided this stops when a boundary or

5 optimum is reach, namely when

$$\frac{\partial U_r(x_i)}{\partial x_i} = \sum_{j \in r} p_j(y_j).$$

Conclusion

Congestion avoidance within aggregate channels has been described. Although

10 specific embodiments have been illustrated and described herein, it will be appreciated by

those of ordinary skill in the art that any arrangement which is calculated to achieve the

same purpose may be substituted for the specific embodiments shown. This application is

intended to cover any adaptations or variations of the present invention. Therefore, it is

manifestly intended that this invention be limited only by the following claims and

15 equivalents thereof.